Computação e Linguística: revisitando a vertente tecnológica da pesquisa em linguística[1]

Agradeço a Fundação de Amparo à Pesquisa do Estado da Bahia – Fapesb por estimular a prática de pesquisa e criar as condições ao desenvolvimento da iniciação científica

Victor Moreira Rocha Brandão[2],
Profª. Drª. Cristiane Namiuti (Orientadora)[3]

Resumo

O presente resumo expandido apresenta o atual estado da pesquisa do projeto de iniciação científica (Fapesb), que se vincula a um projeto maior intitulado "Sintaxe diacrônica em corpus eletrônico: do português pré-clássico às variantes modernas", coordenado pela Profª Cristiane Namiuti e realizado no LAPELINC (Laboratório de Pesquisa em Linguística de *Corpus*). O projeto de IC teve como proposta dar continuidade à investigação do estado da arte sobre as tecnologias de produção de corpora e tratamento de dados para a pesquisa no âmbito das Humanidades Digitais, especificamente da Linguística de Corpus, abrangendo assim as questões referentes a construção, limites e possibilidades da área de um modo que possa ter uma aplicação no método LAPELINC. Durante o segundo ano de iniciação científica, foi possível concluir que o projeto é de suma importância na formação de pesquisadores, evidenciando o quanto a interdisciplinaridade é fundamental para a construção e o estudo de *corpora* linguísticos.

Palavras-Chave: Humanidades Digitais. LAPELINC. Linguística de *Corpus*.

Computing and Linguistics: revisiting the technological side of the research on linguistics

Abstract

The present expanded abstract presents the current state of research of the scientific initiation project (Fapesb), which is linked to a larger project entitled "Sintaxe diacrônica

[1] Pesquisa desenvolvida e financiada por intermédio da Fundação de Amparo à Pesquisa do Estado da Bahia - Fapesb da Universidade Estadual do Sudoeste da Bahia – UESB.

[2] Discente do curso de bacharelado em ciência da computação e bolsista IC da Fapesb– 2020/2021. Endereço Profissional: Universidade Estadual do Sudoeste da Bahia - UESB, estrada do Bem-querer, km. 4, campus universitário, 45083900 – Vitória da Conquista, BA – Brasil – caixa postal – 95. E-mail para contato: victor.brandao88@gmail.com

[3] Doutora em Linguística pela Universidade Estadual de Campinas – UNICAMP, com estágio sanduíche na Universidade de Lisboa, Portugal. Professora Titular do Departamento de Estudos Linguísticos e Literários na Universidade Estadual do Sudoeste da Bahia – DELL/UESB. Professora do Programa da Pós-Graduação em Linguística – PPGLIN/UESB. Endereço Profissional: Universidade Estadual do Sudoeste da Bahia - UESB, estrada do Bem-querer, km. 4, campus universitário, 45083900 – Vitória da Conquista, BA – Brasil – caixa postal – 95. E-mail para contato: cristianenamiuti@uesb.edu.br

em corpus eletrônico: do português pré-clássico às variants modernas", coordinated by the teacher Cristiane Namiuti and carried out at LAPELINC (Laboratório de Pesquisa em Linguística de Corpus). The CI project had the purpose of continuing the investigation on the state of art about the technologies of corpora production and data treatment for the research on the Digital Humanities scope, specifically on Linguistics Corpus, thus covering the questions concerning the construction, limits and possibilities of the area in a way that can be applied on the LAPELINC method. During the second year of the scientific initiation, it was possible to conclude that the project is of great importance on the formation of researchers, highlighting how much the interdisciplinarity is fundamental on the studies of linguistics *corpora*.

Keywords: *Corpus* Linguistics. Digital Humanities. LAPELINC.

Introdução

Segundo Namiuti e Santos (2016), as fontes documentais que fundamentaram os estudos em humanidades durante diversos períodos da história possuem materialidade restrita e ligada ao tempo e espaço. Conforme dito pelos autores, o suporte material dessas fontes é limitado e o manuseio necessita de muito cuidado. O projeto de Iniciação Científica (IC) vincula-se ao projeto temático “Sintaxe diacrônica em corpus eletrônico: do português pré-clássico às variantes modernas” coordenado pela Professora Cristiane Namiuti e que tem seu desenvolvimento no Laboratório de Pesquisa em Linguística de Corpus (LAPELINC), visando contribuir na construção de um corpus digital advindos dos documentos notariais manuscritos guardados em arquivos da região do Sudoeste Baiano: o Corpus DOViC (Corpus de Documentos Oitocentistas de Vitória da Conquista e região).

Para tanto, o presente projeto de IC, orientado pela Profª. Drª. Cristiane Namiuti (UESB/DELL) e co-orientado pelo Prof. Dr. Jorge Viana Santos (UESB/DELL), visou continuar investigando o estado da arte da produção de corpora e tratamento de dados para pesquisa de humanidades digitais, mostrando o quanto a tecnologia é importante no estudo de linguística e como ambas as áreas andam lado a lado. Neste âmbito, o projeto “Computação e Linguística: revisitando a vertente tecnológica da pesquisa em linguística” dá continuidade ao diálogo tão essencial entre as duas áreas: Computação e Linguística.

Nesse presente ano de atividades realizadas de forma híbrida devido ao processo de retomada presencial após a pandemia de COVID-19, foi possível a participação, por meio da plataforma *Google Meet*, em grupos de estudo e pesquisa; reuniões de orientação; participação como ouvinte em minicursos e eventos acadêmicos, possibilitando assim uma grande aquisição de conhecimentos. Também foram feitos testes com ferramentas do LAPELINC além da participação na construção

do Dossiê de Observações Pertinentes (DOP) (SANTOS; NAMIUTI 2019) de documentos notariais antigos, guardados no arquivo municipal de Rio de Contas para fazerem parte do corpus DOViC.

Material e Métodos

Durante o período de seis meses da bolsa, por meio de reuniões da plataforma *Google Meet*, foi possível atingir algumas das metas do plano de trabalho. As metas previam a leitura do referencial teórico e metodológico; prática em pesquisa; produção de relatórios semestral e final Fapesb (abordando atividades desenvolvidas no presente ano de bolsa).

Resultados e Discussão

No que diz respeito à leitura do referencial teórico, durante a duração da Iniciação Científica foi realizada a leitura e discussão de produções de diversos autores, podendo citar as produções dos pesquisadores do LAPELINC, Namiuti (2011), além de textos de Namiuti, Faria e Souza (2017), Kepler e Finger (2006), Souza (2017), Araripe (2010).

A partir dos estudos realizados, foi possível compreender como as fontes documentais ganharam novos suportes com o advento da tecnologia, uma vez que os documentos físicos possuem materialidade restrita, ligada ao tempo e espaço e, portanto necessitam de extremo cuidado por serem frágeis e terem suporte material limitado. Entretanto, também é necessário que seja garantida a fidedignidade desses documentos quando forem passados para o meio digital, e para solucionar os problemas dessa passagem de Documento Físico (DF) para Documento Digital Texto (DDT) foi criado o método LAPELINC para a construção de corpora eletrônicos anotados e cientificamente controlados.

O método LAPELINC consiste em uma etapa de transposição do DF para o Documento Digital Imagem (DDI), que é feita utilizando da fotografia na mesa cartesiana, uma segunda etapa de transcrição para gerar um DDT, que envolve leitura e transcrição paleográfica e uma terceira etapa de compilação de corpora, em que anotações de edição e estrutura morfossintática são inseridas no DDT em XML. Utilizando o método, é possível criar a visão completa do livro, sendo o mesmo ordenado e com meta-informações, e ter sua edição segura realizada por meio da ferramenta WebSinC após a integração ao banco de dados.

Nesse sentido, participamos de um experimento envolvendo o Lapelinc-Transcriptor, uma ferramenta de transcrição automática envolvendo inteligência artificial que está em desenvolvimento no laboratório no âmbito do doutorado de Bruno Silvero Costa, pesquisador do Lapelinc, sob a orientação do Professor Jorge Viana Santos e co-orientação da Professora Cristiani Namiuti. Esse trabalho gerou uma participação em coautoria no II Congresso Internacional de Paleografia e Diplomática durante a sessão II: Paleografia, texto e contexto: o estudo da ortografia e a construção da história.

Conclusões

Observa-se então, a importância que o projeto possui nos estudos de corpora por conta da interdisciplinaridade dos conteúdos. O período de vigência da bolsa pode ser considerado produtivo, já que foi possível cumprir as etapas previstas para os 6 meses de duração, além de ter sido possível aprender na prática sobre o método LAPELINC, os meios e ferramentas necessárias para aplicá-lo e a possibilidade de manusear documentos antigos através do Dossiê de Observações Pertinentes.

Referências Bibliográficas

ARARIPE, Leonel Figueiredo de Alencar. **Aelius: uma ferramenta para anotação automática de corpora usando o NLTK.** In: ENCONTRO DE LINGUÍSTICA DE CORPUS, 9, 2010, Porto Alegre. Anais ... Porto Alegre: PUCRS, 2010. 8P.

KEPLER F. N, FINGER M. (2006) **Part-of-Speech Tagging of Portuguese Based on Variable Lenght Markov Chains.** In: Vieira R., Quaresma P., Nunes M..G.V., Mamede N.J., Oliveira C., Dias M.C. (eds) Computational Processing of the Portuguese Language. PROPOR 2006. Lecture Notes in Computer Science, vol 3960. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11751984_32

NAMIUTI, Cristiane. **Uma reflexão gerativista sobre a dimensão dinâmica de mudanças sintáticas na diacronia do português.** In: XVI Congresso Internacional da ALFAL. Alcalá de Henares-Espanha, 8 de junho de 2011.

NAMIUTI, C., FARIA, P. P., SOUZA, L.F.C. (2017). **eASSIGNER: UMA PROPOSTA DE AUTOMAÇÃO DAS EDIÇÕES DE ANOTAÇÕES XML DO Edictor.** A cor das letras, 17(1), 67-76. DOI: 10.13102/cl.v17i1.

SANTOS, Jorge Viana; NAMIUTI, Cristiane. **O futuro das Humanidades Digitais é o passado.** In: Carrilho, E, Martins, AM, Pereira S; Silvestre JP, Organizadores. Estudos Linguísticos e Filológicos Oferecidos a Ivo Castro. Lisboa: Centro De Linguística Da Universidade de Lisboa; 2019. http://hdl.handle.net/10451/39619. ISBN 978-989-98666-3-8.

SOUZA, Luis Fernando Cardeal de. **eAssigner: concepção e modelagem de software para a automatização de anotações filológico-linguisticas em corpora eletrônicos usando XML.** Agosto de 2017. 109 folhas. (Programa de Pós-Graduação em Linguística – PPGLIN) – Universidade Estadual do Sudoeste da Bahia, Agosto de 2017.